文章编号: 1001-1595(2002)S0-0099-04 中图分类号: P208 文献标识码: A

GIS 数字地图质量子幅抽样方案的探讨

刘 春[1,2], 刘大杰[1], 史文中[2]
(1. 同济大学 测量与国土信息工程系, 上海 200092; 2. 香港理工大学 土地测量与地理资讯系, 香港)

Study of Quality Sampling Inspection with Seed-map to Digital Products in GIS

LIU Chun[1,2], LIU Da-jie[1], SHI Wen-zhong[2]
(1. *Department of Surveying and Geo-informatics, Tongji University, Shanghai* 200092, *China*; 2. *Department of Land-surveying and Geo-informatics, The Hong Kong Polytechnic University, Hong Kong, China*)

Abstract: On the basis of the fundamental of the quality sampling inspection and the sampling inspection schemes to digital product, brings forward the method of using the 1/ *m* seed-map as a unit to conduct the sampling inspection, and the OC curve is given to analyse the feasibility of the seed-map sampling. Meanwhile, the existed sampling inspection schemes can be adjusted to satisfy the need of sampling inspection with seed-map.

Key words: sampling inspection; digital products; GIS

摘 要: 在对数字地图产品实行抽样检验的基本原理和抽样方案的基础上, 对数字地图提出了一种新的 1/ *m* 子幅质量抽样检验方法, 并对其原理与方法进行了探讨, 同时以特征曲线(OC 曲线)分析了这种方法的可行性, 这种新的 1/ *m* 子幅抽样检验方法也可以通过调整已有的整幅抽样方案而获得。

关键词: 抽样检验; 数字地图; GIS

1 前 言

对 GIS 来说, 数据质量的优劣将直接影响 GIS 应用分析结果的可靠程度和应用目标的实现, 因为数字系统比模拟系统处理数据更为精确。但它们的所有数据质量仍然取决于它们采用的数据源。

随着更多的数字产品进入测绘市场和 GIS 应用领域, 对其产品质量进行科学的检验和评价已迫在眉睫。抽样检验可以对产品的质量管理提出可靠的信息, 是质量控制的基本手段。因此, 探讨 GIS 产品的抽样检验方法对 GIS 的质量控制也是十分重要的。文献[1] 介绍了地理数据统计质量控制的概念, 并探讨了采用 OC 曲线确定在一定质量水平下的抽样方案。我们在文献[2] 中对我国数字地图产品的抽样检验方法作了分析, 认为在测绘数据产品中以检验批的 30%和 10%进行抽样检验基本可以满足质量检验的要求, 但考虑到百分比抽样检验在理论上的缺陷, 建议改为采用挑选型抽样检验的方法。本文则探讨了对 GIS 数字产品采用 1/ *m* 子幅为抽样单位的抽样检验方法, 并通过接受概率的计算和 OC 曲线的绘制

收稿日期: 2001-12-06; 修回日期: 2002-01-30
基金项目: 国家自然科学基金资助项目(40171078)
作者简介: 刘 春(1973-), 男, 江苏海门人, 讲师, 博士后, 现从事地理信息系统理论、应用及数据质量领域的研究。

来分析采用这种方法的可行性。

2 以 1/*m* 子幅为抽样单位的接受概率

对一批产品进行抽样检验时，其检验结果可能被接受，也可能被拒绝。通常称可能被接受的概率为接受概率。一批产品的接受概率与批不合格品率 p 密切相关，故常将接受概率记为 $L(p)$。对于批量为 N 的被检验产品，一次抽样检验方案用(n, c)表示，其中 n 表示抽样的样本大小，c 表示合格判定数，当检验后查出不合格品数为 i，则判定规则是：当 $i \leqslant c$ 时接受此批产品；当 $i > c$ 时则拒绝此批产品。

数字地形图每一图幅包括的内容比较丰富，所以以一个图幅为抽样检验单位，抽样时间和费用就相对比较大。如果把一个图幅为抽样单位改为选取其中的 1/m 子幅作为抽样单位来实施抽样检查，那么抽样时间和抽样费用显然就小得多。然而采用 1/m 子幅来代替整图幅实施抽样检查时，抽样方案将由于抽样接受概率的变化而发生变化。本文把以 1/m 子幅为抽样单位进行抽样检验的方法称为子幅抽样，子幅抽样比较适合数字地图数据的质量检验，对子幅抽样方案的探讨也有助于在GIS数字地图检查中推广质量抽样检验以提高数字产品的质量。

在抽样检查中，以 1/m 子幅代替整幅图作为抽样单位进行抽样检查，一个完整图幅或 1/m 子幅检查是否合格，是根据有关标准的规定，以检查单位中含缺陷数的多少来判断。这里的缺陷是指属性数据质量与标准质量的不符，按照不符合的程度可分为轻缺陷、重缺陷和严重缺陷，缺陷数就是所有检验数据中包含的缺陷总数。而用于判断整幅图幅不合格的缺陷数判定值与判断 1/m 子幅不合格的缺陷数判定标准是不一样的，后者的判定值一般根据前者的判定值进行适当调整，一般是调整判定值，以使得采用 1/m 子幅来检查整幅图的质量更合理。整幅抽样和子幅抽样的合格判定值经过适当调整以后，两者对数据的检验精度应该是等价的，但是两者由于检验数据的数量不同因而检验费用也完全不同。

现假设将一幅图分成 m 子幅，抽取其中一子幅并判断为不合格的事件设为 B，其概率为 $P(B)$。设 j ($j=1, 2, \cdots, m$) 为一幅图分成 m 份后，其中有 j 个子幅为不合格的事件。对于一个不合格的图幅，在没有关于其不合理的先验分布信息时，其缺陷数在这幅图中的位置分布可以认为是随机的。由于 1/m 子幅检查标准不同于整图幅的检查标准，且不合格图中缺陷数在图中的分布是随机的，所以有

$$P(j)=\frac{1}{m} \tag{1}$$

所以当一图幅中有 j 个子幅不合格时，抽取任意一子幅，该子幅不合格的概率为

$$P(B|j)=\frac{C_j^1}{C_m^1}=\frac{j}{m} \tag{2}$$

容易理解，在抽样检查中以 1/m 图幅的子幅代替整幅图作为抽样单位进行抽样检查时，一幅不合格的图以 1/m 图幅的子幅不合格来判断整图幅为不合格的概率是相等的，且为

$$P_{1/m}=P(B)=\sum_{j=1}^{m}P(B|j)P(j)=\frac{1}{m}\left(\frac{1}{m}+\frac{2}{m}+\cdots+\frac{m}{m}\right)=\frac{m+1}{2m} \tag{3}$$

因此，一个不合格的图幅，以 1/m 图幅为单位检查，认为整图幅是合格的概率 $P'_{1/m}$ 为

$$P'_{1/m}=1-P_{1/m}=1-\frac{m+1}{2m}=\frac{m-1}{2m} \tag{4}$$

显然，当取 $m=1$，即不分子幅时 $P'_{1/m}=0$，即检查不合格图幅被认为是合格的概率为 0。

检验批量为 N 幅图的产品中不合格品数为 $D=Np$ (p 为不合格品率)，设以 H_i[$i=0, 1, 2, \cdots, \min(D, n)$] 表示事件“随机地抽取 n 个样本，其中恰有 i 个样本为不合格”，所以 H_0, H_1, H_2, $\cdots$, $H_{\min(D, n)}$ 是检验批抽样结果的一个划分，且以 A 表示事件“数字地图检验批被接受”，其概率 $P(A)$ 也就是检验批被接受的接受概率 $L(p)$。H_i 是一个服从超几何分布的随机度量，其概率为

$$P(H_i)=\frac{C_D^i C_{N-D}^{n-i}}{C_N^n} \tag{5}$$

其中，C_N^n 表示从 N 中取 n 个产品组合数。

当 N 较大，$\frac{n}{N} \leqslant 0.10$ 时，可以按照二项分布计算 $P(H_i)$，即有

$$P(H_i)=C_n^i p^i (1-p)^{n-i} \tag{6}$$

由此可知，采用一次抽样方案，n 个样本的抽取是相互独立的，其合格判定数为 c，当以 1/m 子幅代替整幅图进行检验时，在有 i 个子幅不合格的情况下，检验认为这批数字地图产品为合格的接受概率为

$$L(p)=\sum_{i=0}^{c}C_n^i\left(\frac{m+1}{2m}p\right)^i\left(1-\frac{m+1}{2m}p\right)^{n-i} \quad (7)$$

由式(7)知，接受概率不仅与 n，c 有关，还与 m 有关，将这种抽样方案表示为(n，c，m)。当 m =1 时，(n，c，1)也就是以整幅图为单位的抽样方案(n，c)，而式(7)变为

$$L(p)=\sum_{i=0}^{c}C_n^i p^i(1-p)^{n-i} \quad (8)$$

式(8)为相应的接受概率，该式与文献[2]所得到的数字地图抽样接受概率公式一致。

3　以 1/m 子幅为抽样单位的 OC 曲线

接受概率 $L(p)$是不合格品率 p 的递减函数，通常将 $L(p)$与 p 之间的函数关系在直角坐标系中，表示成抽样检验特性曲线(Operating Characteristic Curve，简称为 OC 曲线)。不同的抽样检验方案，对应着不同的 OC 曲线。OC 曲线具有区分抽样方案拒绝能力的作用，是建立和选择抽样检验方案的一种依据，所以 OC 曲线反映了抽样特性，这种抽样特性也就是抽样方案区分数据质量好坏的能力，以及是否以高概率接受质量好的数据且以高概率拒绝质量差的数据。

根据式(7)，确定抽样方案(n，c，m)以后，OC 曲线是惟一确定的，现以 OC 曲线分析以1/m子幅为抽样单位的抽样检验特性。

1．n 和 c 不变，m 对 OC 曲线的影响。如取 n=50，c=1；m=1，2，4，8。这就构成 4 个抽样方案(50，1，1)，(50，1，2)，(50，1，4)，(50，1，8)。由这些抽样方案的OC 曲线，可以看出，m 越大，OC 曲线越平缓，其区分图幅数据质量好坏的能力减弱，因而抽样方案越宽松，但是从图上也可以看出，m 的变化对OC 曲线平缓的变化影响比较小。可以认为，m 增大能导致抽样宽松，但其影响不大。

2．c 不同时，m 对 OC 曲线的影响。对于一个抽样方案，c 作为合格判定数，来判断批是否被接受，c 的大小决定了抽样的严格与否。由图1(a)，1(b)，1(c)可发现，m 的增大使得图 1 中的OC 曲线向上移，也就是接受概率增大。

3．由图 1 的 OC 曲线可以看到，影响抽样特性的主要是 n，c 2 个量，它们基本决定了 OC 曲线的形状，也即基本决定了抽样方案区分好坏的能力。采用 1/m 图幅为抽样单位进行抽样检查时，一方面可以节约抽样时间和费用；另一方面对于由 n，c 确定的抽样特性影响不大，尤其是在合格判定数很小时。

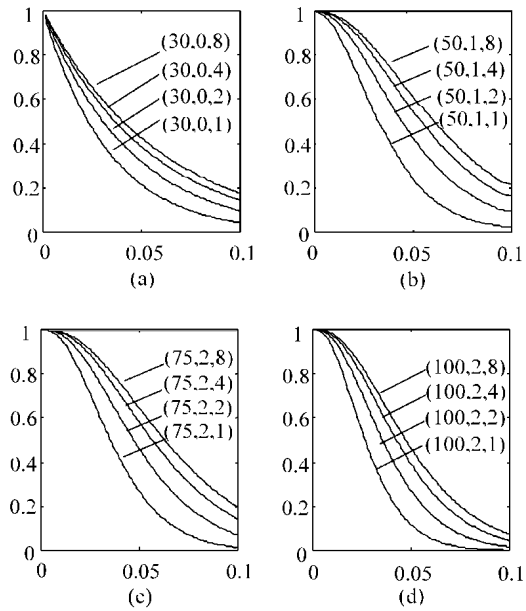

图 1　OC 曲线

Fig. 1　OC Curve

4　子幅抽样方案

挑选型抽样方案的特点是对判为接受的批直接通过，对判为拒绝的批必须经过 100%的检查，将其中不合格品挑出来，换成合格的(或修复为合格品)然后再接受。国家标准 GB/T13546 规定了挑选计数一次抽样方案及实施程序。文献[2]对 GIS 数字地图产品的抽样检验方案选择作了探讨，并说明了采用挑选型计数抽样方案较适合目前GIS 数字地图产品质量抽样检验的特点。

挑选型计数一次抽样方案有以下 2 种，计数挑选一次极限质量抽样和计数挑选型一次检验后平均不合格品率上限抽样。现以计数挑选一次极限质量抽样为例说明 1/m 图幅为抽样单位的抽样检验方案的选择。

设对于一批产品，其批量为 N，假定此种产品的平均不合格品率为 p，如果采用以整幅图为抽样单位的一次抽样方案(n，c)，当 N 较大，n/N ≤0.10，p≤0.10 时，接受概率$L(p)$可按式(7)计算。按照均值的概念，挑选型一次抽样方案的平均检验件数为

$$\bar{I}=L(p)n+[1-L(p)]N=n+(N-n)[1-L(p)] \quad (9)$$

如果采用子幅抽样方案(n'，c'，m)，按式(8)计算其接受概率 $L_m(p)$，则相应的平均检验件数为

$$\bar{I}_m = n + (N-n)[1 - L_m(p)] \quad (10)$$

一般来说, 当取 $n' = n$, $c' = c$ 时, 有 $L_m(p) \geqslant L(p)$, $\bar{I}_m \geqslant \bar{I}$ 即采用子幅抽样检查会使抽样宽松。为了避免产生这种抽样宽松的现象, 可以采用取 $c' = c$, 而使 c' 满足 $\bar{I}_m = \bar{I}$ 的要求。如果规定一个极限不合格品率 p_t, 这个极限不合格品率以整图幅为抽样单位的抽样方案的接受概率 $L(p_t)$ 可按式(8)计算, 又采用子幅抽样方案 (n', c, m), 由式(7)计算其接受概率 $L_m(p_t)$, 则由

$$\bar{I} = n + (N-n)[1 - L(p_t)] = n'(N-n'')[1 - L_m(p_t)] = \bar{I}_m$$

可得

$$n' = \frac{N[L_m(p_t) - L(p_t)] + nL(p_t)}{L_m(p_t)} \quad (11)$$

也就是说, 子幅抽样方案应为 (n', c, m), 其中 n' 由式(11)求得。表 1 给出了挑选型一次极限质量抽样的 2 类抽样方法的抽样方案。

表 1 2 类抽样方法的挑选型一次极限质量抽样方案

Tab. 1 Sampling schemes of the one time limiting of count selection for the two sampling methods

极限质量 p_t/(%)	批量大小 N	整图幅抽样抽样方案 (n, c)	子幅抽样抽样方案 $(n', c, m=4)$
2.00	90	全检	全检
	150	(80, 0)	(92, 0)
	280	(89, 0)	(110, 0)
5.00	50	全检	全检
	90	(33, 0)	(48, 0)
	150	(37, 0)	(62, 0)
	280	(42, 0)	(81, 0)

由表 1 可以看到, n' 略大于 n。显然, 随着 n' 的增大, 抽样变严格, 这样就能消除了由于采用子幅抽样检查引起的抽样宽松, 而且绘制的同一极限质量 2 种抽样方案的 OC 曲线, 发现差别不大, 因而说明这种调整是合适的。

还可以看到, 对于检验批 N, 采用子幅抽样方案 (n, c, m) 的工作量是采用整幅抽样方案 (n, c) 工作量的 $1/m$, 而抽样方案 (n', c, m) 的工作量与抽样方案 (n, c) 工作量之比为

$$\frac{n'}{mn} = \frac{1}{m}\left\{1 + \frac{[L_m(p_t) - L(p_t)](N-n)}{nL_m(p_t)}\right\} \quad (12)$$

5 结 论

本文在探讨 GIS 数字地图质量抽样检验所涉及的基本原理与方法的基础上, 提出了一种新的采用 $1/m$ 子幅作为抽样单位对 GIS 数字地图进行抽样检验的方法, 通过以上的分析, 可以得出以下结论:

1. 采用 $1/m$ 子幅代替整个图幅实施抽样检验可以较大的降低抽样时间和费用。这里选取的 $1/m$ 子幅可以是一整图幅的 $1/m$ 子幅, 也可以是以几幅图为一个整体的 $1/m$ 子幅。

2. 采用 $1/m$ 子幅作为抽样单位进行抽样检验时, 若取 $n' = n$, $c' = c$, 会引起抽样的宽松, 但这种宽松与合格判定数 c 有关, 合格判定数 c 越小, 抽样宽松的影响就越小。

3. 选择挑选型计数抽样方法, 采用 $1/m$ 图幅进行抽样时, 可以按式(11)确定 n' 值, 并保持合格判定数 c 不变, 从而通过这种抽样加严来弥补由于采用 $1/m$ 图幅抽样引起的抽样宽松。

4. 本文所提出的子幅抽样方法一方面有其严密的理论基础, 另一方面这种方法确实能够在大大降低原来检验费用的同时而获得近似的抽样检验精度, 因而有实用价值, 值得推广使用。

值得说明的是, 本文关于以 $1/m$ 子幅作为抽样单位的讨论还是初步的, 还有不少具体问题尚待进一步探讨。

参考文献:

[1] CASPARY W, JOOS G. Statistical Quality Control of Geodata[A]. Precedings of the International Symposium on Spatial Data Quality'99[C]. Hong Kong: [s. n.], 1999.

[2] LIU Da-jie, LIU Chun. Sampling Inspection to Digital Products in GIS[A]. Precedings of the International Symposium on Spatial Data Quality'99[C]. Hong Kong: [s. n.], 1999.

[3] LIU Da-jie, *et al*. Accuracy Analysis and Quality Control of Spatial Data in GIS[M]. Shanghai: Shanghai Press of Science and Technique, 1999. (in Chinese)

[4] YU Shang-qi. Sampling Inspection and Quality Control [M]. Beijing: Beijing University Press, 1991. (in Chinese)

[5] MA Yi-lin, YU Zhen-fan, YU Shang-qi. Sampling Inspection for Product Quality[M]. Beijing: China Press of Standard, 1997. (in Chinese)